Large Language Models Enhanced Covariate-adjusted Response-adaptive Randomization Design

Yanping Li
Division of Biostatistics
University of California, Berkeley
yanpingli@berkeley.edu

Xinrui Ruan
Division of Biostatistics
University of California, Berkeley
xinruiruan@berkeley.edu

Jingshen Wang*
Division of Biostatistics
University of California, Berkeley
jingshenwang@berkeley.edu

Waverly Wei*
Department of Data Sciences and Operations
University of Southern California
waverly@marshall.usc.edu

Abstract

Covariate-adjusted response-adaptive randomization (CARA) designs can improve statistical efficiency and participant welfare in randomized experiments by learning from accrued data and dynamically updating treatment allocation. In practice, two issues limit their impact. Many experiments are sample-size constrained because enrollment is difficult and trials are expensive. At the same time, studies increasingly collect rich pre-treatment information, including unstructured data such as text, images, and clinical notes, yet most CARA designs rely on a small set of structured covariates when computing allocation probabilities. Recent advances in large-scale pre-trained large language models (LLMs) offer a new opportunity to extract useful signals from unstructured inputs and external knowledge. However, few-shot LLM predictions are not ordinary fixed fitted predictions because prompt demonstrations sampled from accrued trial data create data-dependent randomness and can induce correlation across predictions. If used naively to drive allocation updates, AI-generated signals may weaken CARA's ability to improve power and participant welfare. We propose a CARA design that integrates few-shot LLM-predicted outcomes through resampling-based aggregation, calibration, and effective residual variances to guide both adaptive treatment allocation and downstream treatment effect estimation. Theoretical results and simulation studies show that, when calibrated few-shot LLM auxiliary signals reduce residual variation, coupling an LLM-enhanced estimator with an adaptive allocation rule can improve statistical efficiency and participant welfare.

1 Introduction

Covariate-adjusted response-adaptive randomization (CARA) designs learn from accrued trial data and update treatment assignment probabilities over stages, with the dual aims of improving statistical efficiency and promoting ethical or welfare-oriented goals [Hu and Rosenberger, 2006, Rosenberger et al., 2001, Hu et al., 2015, Rosenberger and Lachin, 2015]. These aims are particularly compelling in modern randomized experiments that are *sample-size constrained* because recruitment is difficult and trials are expensive, yet *data rich at baseline* because pre-treatment records increasingly include structured variables and unstructured inputs such as free-text surveys, clinical notes, and images. In principle, CARA can use accumulating trial information to improve both power and participant

*Jingshen Wang and Waverly Wei are co-corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

welfare. In practice, however, CARA's empirical gains are often under-explored in these regimes [Robertson et al., 2023, Offer-Westort et al., 2021].

Two practical issues drive this gap. First, when interim sample sizes are small, stagewise estimates are noisy, which can make allocation updates unstable [Hu and Rosenberger, 2006, Zhang et al., 2007, Hu and Hu, 2012]. Second, most CARA procedures rely on a small set of structured covariates when computing allocation probabilities. High-dimensional and unstructured baseline inputs are harder to use during enrollment because sequentially fitted prediction models can overfit and naive plug-in predictions can inject noise or bias. As a result, prognostic information in unstructured baseline records is often ignored, even when it could improve both design and analysis.

Recent advances in large-scale pretrained language models (LLMs) offer a way to extract auxiliary prognostic signals from unstructured baseline inputs in sample-size constrained trials. Motivated by evidence that LLM-based predictions can increase power at the analysis stage [Ruan et al., 2025], a natural next step is to use few-shot LLM signals to guide adaptive allocation. Unlike analysis-stage LLM adjustment for fixed randomized experiments, our setting uses few-shot LLM auxiliary signals inside a CARA design to update future treatment allocation probabilities, requiring predictable randomization and inference under adaptive allocation. Few-shot LLM prediction differs from standard supervised prediction because it is not a fixed fitted function. It conditions on demonstration examples sampled from accrued trial data, which can make predictions data-dependent and correlated across subjects. This stochastic layer is delicate in CARA designs, where allocation probabilities are updated sequentially and validity relies on predictable randomization and martingale arguments. LLM-generated predictions may also contain systematic errors or subgroup miscalibration, so naive use can misdirect assignments and weaken efficiency or welfare.

This manuscript studies how to integrate arm-specific few-shot LLM auxiliary signals into CARA while preserving predictable randomization and valid inference. The design treats LLM outputs as auxiliary signals rather than plug-in outcomes. It aggregates prompt-level predictions through a resampling-based procedure inspired by classical U-statistic theory, calibrates the aggregated signal within strata, and uses the resulting effective residual variances to update allocation probabilities. For efficiency maximization, this yields an LLM-enhanced Neyman-type allocation. For welfare improvement, it yields a constrained allocation rule that trades off participant welfare and statistical precision. The final estimator uses design-known propensities and cross-fitting to support inference under predictable randomization [Tsiatis, 2006, Hahn, 1998, Hirano et al., 2003, Ma and Wang, 2020, Bibaut and Kallus, 2025]. Our work is complementary to adaptive Neyman allocation, batch-adaptive experimentation, Neyman-regret, and semiparametric inference for adaptive experiments [Hahn et al., 2011, Dai et al., 2023, Zhao, 2023, Neopane et al., 2025], whereas our focus is on resampled few-shot LLM auxiliary signals that enter both CARA updates and final inference.

Our contributions are summarized as follows:

1. We introduce a CARA design procedure that uses arm-specific few-shot LLM auxiliary predictions to guide adaptive allocation. The procedure aggregates data-dependent and prompt-randomized outputs, calibrates the resulting signals within strata, converts them into effective residual variances, and updates allocation probabilities under efficiency and welfare objectives.
2. We develop an average treatment effect estimator that uses design-known propensities, cross-fitting, and calibrated LLM signals through residual-variance reduction rather than treating LLM outputs as ground-truth counterfactual outcomes. We establish allocation consistency and asymptotic normality by extending CARA martingale arguments to include the prompt-aggregation layer.
3. We characterize multi-layer efficiency gains from calibrated LLM adjustment, LLM-guided allocation, and few-shot aggregation when it improves the auxiliary signal. Theory and simulations show that these gains can improve efficiency and participant welfare when calibrated auxiliary signals reduce residual variation.

2 CARA: Problem Setup

In this section, we develop a CARA framework that uses few-shot LLM auxiliary predictions from the full baseline record to stabilize allocation updates and improve final-stage inference.

2.1 CARA enrollment

The experiment has $S < \infty$ stages. In stage s, n_s individuals are enrolled, with cumulative sample size $N_s = \sum_{r=1}^{s} n_r$ and total sample size $N = N_S$. Individuals are indexed by enrollment order $i = 1, \ldots, N$, and $S_i = s$ if $N_{s-1} < i \leq N_s$, with $N_0 = 0$.

Following the Neyman-Rubin potential outcomes framework [Neyman, 1923, Rubin, 1974], let $T_i \in \{0, 1\}$ denote treatment assignment and let $Y_i(t)$ denote the potential outcome under arm $t \in \{0, 1\}$. The observed outcome is $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$. The average treatment effect is $\tau = \mathbb{E}\{Y(1) - Y(0)\}$, and the stratum-specific effect is $\tau(z) = \mathbb{E}\{Y(1) - Y(0) \mid Z = z\}$. Each subject has baseline record $W_i = (Z_i, X_i, U_i)$, where $Z_i \in \{1, \ldots, J\}$ is a low-dimensional stratification variable, $X_i \in \mathbb{R}^d$ contains structured covariates, and U_i contains high-dimensional or unstructured inputs such as clinical notes, images, or missing-heavy records.

2.2 CARA allocation rules

Let $\mathcal{G}_i$ denote the information revealed up to individual i, including baseline records, assignments, and all observed outcomes. Outcomes may be delayed. Let $\mathcal{H}_s = \mathcal{G}_{N_s}$ denote the history available at the end of stage s.

A CARA design specifies stratum-specific allocation functions $\boldsymbol{e} = \{e_s(\cdot)\}_{s=1}^{S}$, where $e_s : \{1, \ldots, J\} \to [\delta, 1 - \delta]$ and $\delta \in (0, 1/2)$ enforces positivity. At the beginning of stage s, the design computes $e_s(\cdot)$ from $\mathcal{H}_{s-1}$ and assigns each newly enrolled subject according to

$$\Pr(T_i = 1 \mid \mathcal{G}_{i-1}, Z_i = z) = e_s(z), \quad z \in \{1, \ldots, J\}. \tag{1}$$

2.3 LLM-based counterfactual predictions

To use the full baseline record W_i, we construct prompt-level few-shot predictions $Y_i^{\dagger}(t; \mathcal{S}_{i,t}^{(b)})$, $b = 1, \ldots, B$, where $\mathcal{S}_{i,t}^{(b)}$ is a demonstration set sampled from accrued trial data using only information available before prediction. We aggregate these predictions as $Y_i^{\dagger}(t) = B^{-1} \sum_{b=1}^{B} Y_i^{\dagger}(t; \mathcal{S}_{i,t}^{(b)})$. The design uses this aggregated signal only after centering and calibration.

For each arm t, define $\mu_t^{\dagger}(z) = \mathbb{E}\{Y^{\dagger}(t) \mid Z = z\}$ and $\widetilde{Y}_i^{\dagger}(t) = Y_i^{\dagger}(t) - \mu_t^{\dagger}(Z_i)$. Within stratum z, define the projection slope

$$\omega_t(z) = \frac{\mathbb{C}\mathrm{ov}\{Y(t), \widetilde{Y}^{\dagger}(t) \mid Z = z\}}{\mathbb{V}\mathrm{ar}\{\widetilde{Y}^{\dagger}(t) \mid Z = z\}},$$

with $\omega_t(z) = 0$ if the denominator is zero. The effective residual variance is

$$\tilde{v}_t(z) = \mathbb{V}\mathrm{ar}\{Y(t) - \omega_t(z)\widetilde{Y}^{\dagger}(t) \mid Z = z\} = v_t(z)\{1 - \rho_t^2(z)\}, \tag{2}$$

where $v_t(z) = \mathbb{V}\mathrm{ar}\{Y(t) \mid Z = z\}$ and $\rho_t(z) = \mathbb{C}\mathrm{orr}\{Y(t), Y^{\dagger}(t) \mid Z = z\}$. Thus, calibrated auxiliary predictions reduce the residual uncertainty driving allocation updates when they are predictive within arm-stratum cells.

2.4 Two design objectives

A CARA policy $\boldsymbol{e}$ shapes treatment assignments across stages and strata. We use two oracle objectives to motivate the proposed updating rules.

Objective I. Efficiency maximization. The first objective chooses allocation probabilities to minimize the allocation-sensitive part of the large-sample variance of the final ATE estimator. Let this criterion be $\mathcal{V}(\boldsymbol{e})$. The oracle design solves

$$\min_{\boldsymbol{e}} \ \mathcal{V}(\boldsymbol{e}) \quad \text{s.t.} \quad e_s(z) \in [\delta, 1 - \delta], \quad s \leq S, \quad z \in \{1, \ldots, J\}. \tag{3}$$

Appendix C shows that the solution decomposes across strata and does not depend on the stage index. For each stratum z, the optimal allocation has the LLM-enhanced Neyman form

$$e_{\mathrm{I}}^{*}(z) = \frac{\sqrt{\tilde{v}_1(z)}}{\sqrt{\tilde{v}_1(z)} + \sqrt{\tilde{v}_0(z)}}. \tag{4}$$

This rule replaces raw within-stratum variances with effective residual variances after calibrated auxiliary adjustment.

Objective II. Welfare improvement under an efficiency constraint. Many trials also prioritize participant welfare, such as reducing the probability of failure outcomes. Let $\mathcal{F}$ denote a failure set and define $\mu_t^{\mathrm{F}}(z) = \Pr\{Y(t) \in \mathcal{F} \mid Z = z\}$. The population expected failure rate under policy $\boldsymbol{e}$ is

$$\mathcal{L}_{\mathrm{F}}(\boldsymbol{e}) = \sum_{s=1}^{S} r_s \mathbb{E}\left[e_s(Z)\mu_1^{\mathrm{F}}(Z) + \{1 - e_s(Z)\}\mu_0^{\mathrm{F}}(Z)\right]. \tag{5}$$

The welfare-oriented oracle design solves

$$\min_{\boldsymbol{e}} \ \mathcal{L}_{\mathrm{F}}(\boldsymbol{e}) \quad \text{s.t.} \quad e_s(z) \in [\delta, 1-\delta], \quad s \le S, \quad z \in \{1, \ldots, J\}, \quad \mathcal{V}(\boldsymbol{e}) \le C, \tag{6}$$

where $C > 0$ controls the allowed loss in statistical precision. For binary outcomes with $Y = 1$ indicating success, one may take $\mathcal{F} = \{0\}$, so that $\mu_t^{\mathrm{F}}(z) = 1 - \mathbb{E}\{Y(t) \mid Z = z\}$.

The next section gives an implementable stagewise procedure that estimates these primitives from the evolving history, updates allocation probabilities predictably, and supports valid final inference using realized design-known propensities.

3 Our Proposed Method: LLM-enhanced CARA

In this section, we propose an LLM-enhanced CARA design that uses few-shot LLM auxiliary predictions for adaptive allocation and final inference. The design treats LLM outputs as auxiliary signals rather than plug-in outcomes. Because prompt-level predictions can vary with sampled demonstrations, we aggregate over prompt-sampling randomness in a resampling step inspired by U-statistic symmetrization. We then calibrate the aggregated signal against observed outcomes and convert it into effective residual variances for allocation updates.

Stage 1. Enroll n_1 participants and assign treatment using a pre-specified rule, for example $e_1(z) = 1/2$ for all strata z. For each enrolled subject i, construct arm-specific aggregated few-shot LLM auxiliary signals $Y_i^{\dagger}(0)$ and $Y_i^{\dagger}(1)$ using only information available before prediction, as described in Section 2.3.

Stage $s+1$, for $s = 1, \ldots, S-1$. At the beginning of stage $s+1$, compute the next allocation rule using $\mathcal{H}_s$.

Step 1. Estimation of calibration primitives. Let $\mathcal{O}_s = \{i \le N_s : Y_i \text{ is observed by time } \mathcal{H}_s\}$ and $\mathcal{I}_{t,s}(z) = \{i \in \mathcal{O}_s : T_i = t,\ Z_i = z\}$. Let $\widehat{\mu}_{t,s}(z)$ and $\widehat{\mu}_{t,s}^{\dagger}(z)$ estimate $\mu_t(z) = \mathbb{E}\{Y(t) \mid Z = z\}$ and $\mu_t^{\dagger}(z) = \mathbb{E}\{Y^{\dagger}(t) \mid Z = z\}$, possibly by regression or cross-fitting. Using $\mathcal{I}_{t,s}(z)$, estimate

$$\widehat{\omega}_{t,s}(z) = \frac{\sum_{i \in \mathcal{I}_{t,s}(z)} \{Y_i - \widehat{\mu}_{t,s}(z)\}\{Y_i^{\dagger}(t) - \widehat{\mu}_{t,s}^{\dagger}(z)\}}{\sum_{i \in \mathcal{I}_{t,s}(z)} \{Y_i^{\dagger}(t) - \widehat{\mu}_{t,s}^{\dagger}(z)\}^2},$$

with $\widehat{\omega}_{t,s}(z) = 0$ if the denominator is zero. When $|\mathcal{I}_{t,s}(z)| > 0$, define

$$\widehat{\widetilde{v}}_{t,s}(z) = \frac{1}{|\mathcal{I}_{t,s}(z)|} \sum_{i \in \mathcal{I}_{t,s}(z)} \{Y_i - \widehat{\mu}_{t,s}(z) - \widehat{\omega}_{t,s}(z)(Y_i^{\dagger}(t) - \widehat{\mu}_{t,s}^{\dagger}(z))\}^2.$$

If $|\mathcal{I}_{t,s}(z)| = 0$, keep the previous-stage value or use a pre-specified default.

In small samples or under possible subgroup miscalibration, the auxiliary signal may spuriously appear predictive within an arm–stratum cell. As an optional finite-sample safeguard, one may replace the calibration primitives $\widehat{\omega}_{t,s}(z)$ and $\widehat{\widetilde{v}}_{t,s}(z)$ by validation-gated versions that screen out cells without statistically positive out-of-fold residual-variance reduction; details are given in Appendix A.1. In applications where auxiliary-signal reliability is uncertain, we recommend the validation-gated version as a finite-sample safeguard; the unscreened calibrated version is reported as the main procedure to isolate the target mechanism. Unless otherwise stated, the theoretical development below is written for the main calibrated procedure without this screening step.

When few-shot prompting is used, query the LLM B times with prompt sets $\mathcal{S}_{i,t}^{(b)}$ and aggregate $Y_i^\dagger(t) = B^{-1}\sum_{b=1}^{B} Y_i^\dagger(t;\mathcal{S}_{i,t}^{(b)})$. For final-stage nuisance estimation, use K-fold cross-fitting, with demonstration pools built outside the target fold.

Step 2. Allocation update. Update $e_{s+1}(\cdot)$ according to the selected objective and truncate to $[\delta, 1-\delta]$.

Objective I. Efficiency maximization. For each stratum z, set

$$\widehat{e}_{s+1,\mathrm{I}}(z) = \frac{\sqrt{\widehat{\widehat{v}}_{1,s}(z)}}{\sqrt{\widehat{\widehat{v}}_{1,s}(z)} + \sqrt{\widehat{\widehat{v}}_{0,s}(z)}}. \tag{7}$$

Objective II. Welfare improvement under an efficiency constraint. Let $\widehat{\mu}_{t,s}^{\mathrm{F}}(z)$ estimate $\mu_t^{\mathrm{F}}(z)$, and let $\widehat{p}_s(z)$ be the empirical stratum proportion up to stage s. Choose $\widehat{e}_{s+1,\mathrm{II}}(\cdot)$ by solving

$$\begin{aligned} \min_{e(\cdot)} \quad & \sum_{z=1}^{J} \widehat{p}_s(z)\{e(z)\widehat{\mu}_{1,s}^{\mathrm{F}}(z) + (1-e(z))\widehat{\mu}_{0,s}^{\mathrm{F}}(z)\} \\ \text{s.t.} \quad & e(z) \in [\delta, 1-\delta],\ \forall z, \\ & \sum_{z=1}^{J} \widehat{p}_s(z)\left\{\frac{\widehat{\widehat{v}}_{1,s}(z)}{e(z)} + \frac{\widehat{\widehat{v}}_{0,s}(z)}{1-e(z)}\right\} \le C_s. \end{aligned} \tag{8}$$

By Appendix D, there exists $\lambda_s^* \ge 0$ such that, for each stratum z with $\widehat{p}_s(z) > 0$,

$$\widehat{e}_{s+1,\mathrm{II}}(z) \in \operatorname*{argmin}_{e\in[\delta,1-\delta]} \left[e\{\widehat{\mu}_{1,s}^{\mathrm{F}}(z) - \widehat{\mu}_{0,s}^{\mathrm{F}}(z)\} + \lambda_s^* \left\{\frac{\widehat{\widehat{v}}_{1,s}(z)}{e} + \frac{\widehat{\widehat{v}}_{0,s}(z)}{1-e}\right\}\right], \tag{9}$$

with λ_s^* chosen so that Eq. (8) is tight when $\lambda_s^* > 0$. If the optional validation-gated safeguard is activated, the same formulas are applied after replacing $\widehat{\widehat{v}}_{t,s}(z)$ by $\widehat{\widehat{v}}_{t,s}^{\mathrm{scr}}(z)$. The displayed formulas therefore define the main calibrated allocation update, while the screened version is used only in robustness analyses unless explicitly stated.

Step 3. Treatment assignment and few-shot LLM auxiliary signals for new enrolled subjects. Set $\widehat{e}_{s+1}(\cdot) = \widehat{e}_{s+1,\mathrm{I}}(\cdot)$ under Objective I and $\widehat{e}_{s+1}(\cdot) = \widehat{e}_{s+1,\mathrm{II}}(\cdot)$ under Objective II. For each new subject i with $S_i = s+1$, assign treatment by $\Pr(T_i = 1 \mid \mathcal{G}_{i-1}, Z_i = z) = \widehat{e}_{s+1}(z)$, and compute $Y_i^\dagger(0)$ and $Y_i^\dagger(1)$ from W_i at enrollment.

Statistical inference after stage S. Define $e_i = e_1(Z_i)\mathbb{1}\{S_i = 1\} + \widehat{e}_{S_i}(Z_i)\mathbb{1}\{S_i \ge 2\}$, $e_{i,1} = e_i$, and $e_{i,0} = 1 - e_i$. For each arm t, form

$$\widehat{\varphi}_t(O_i, Y_i^\dagger(t)) = \frac{\mathbb{1}\{T_i = t\}Y_i}{e_{i,t}} + \left(1 - \frac{\mathbb{1}\{T_i = t\}}{e_{i,t}}\right)\left[\widehat{\mu}_t(Z_i) + \widehat{\omega}_t(Z_i)\{Y_i^\dagger(t) - \widehat{\mu}_t^\dagger(Z_i)\}\right].$$

If one also applies the validation-gated safeguard at the analysis stage, $\widehat{\omega}_t(Z_i)$ can be replaced by $\widehat{\omega}_t^{\mathrm{scr}}(Z_i)$. Otherwise the displayed estimator corresponds to the main calibrated procedure. The ATE estimator is

$$\widehat{\tau} = \frac{1}{N}\sum_{i=1}^{N}\{\widehat{\varphi}_1(O_i, Y_i^\dagger(1)) - \widehat{\varphi}_0(O_i, Y_i^\dagger(0))\},$$

with variance estimate

$$\widehat{\mathbb{V}} = \frac{1}{N}\sum_{i=1}^{N}\{\widehat{\varphi}_1(O_i, Y_i^\dagger(1)) - \widehat{\varphi}_0(O_i, Y_i^\dagger(0)) - \widehat{\tau}\}^2.$$

A plug-in estimate of the population failure rate induced by the realized policy is

$$\widehat{\mathcal{L}}_F = \sum_{s=1}^{S} r_s \sum_{z=1}^{J} \widehat{p}_s(z)\{\widehat{e}_s(z)\widehat{\mu}_{1,s}^{\mathrm{F}}(z) + (1-\widehat{e}_s(z))\widehat{\mu}_{0,s}^{\mathrm{F}}(z)\}.$$

4 Efficiency gains and welfare improvements

Our framework couples a calibrated LLM-enhanced estimator with a predictable CARA allocation policy. The estimator uses few-shot LLM auxiliary signals, and the allocation rule uses estimated effective residual variances. This coupling can yield layered efficiency gains under Objective I and welfare benefits under Objective II when the calibrated signal reduces residual variation.

4.1 Three layers of efficiency gains under Objective I

We formalize efficiency comparisons under Objective I using the allocation-relevant oracle criterion $\mathcal{V}(\boldsymbol{e})$. This criterion keeps the terms in the large-sample variance that depend on the allocation policy and omits terms that are invariant to $\boldsymbol{e}$, such as stratum heterogeneity and covariance terms from calibrated augmentation. Appendix B shows that

$$\mathcal{V}(\boldsymbol{e}) = \sum_{s=1}^{S} r_s \mathbb{E}\left[\frac{\tilde{v}_1(Z)}{e_s(Z)} + \frac{\tilde{v}_0(Z)}{1-e_s(Z)}\right], \tag{10}$$

where $r_s = n_s/N$ and $\tilde{v}_t(z) = v_t(z)\{1-\rho_t^2(z)\}$.

Let $\mathcal{V}_{\mathrm{AIPW}}(\boldsymbol{e})$ denote the corresponding allocation-relevant oracle criterion for standard AIPW without auxiliary adjustment. Let $\mathcal{V}^{\mathrm{zs}}$ and $\mathcal{V}^{\mathrm{fs}}$ denote the criteria induced by zero-shot and few-shot auxiliary signals, and let $\boldsymbol{e}_{\mathrm{I}}^{*,\mathrm{zs}}$ and $\boldsymbol{e}_{\mathrm{I}}^{*,\mathrm{fs}}$ denote the corresponding LLM-enhanced Neyman oracle policies. Without the auxiliary signal, the classical Neyman allocation is

$$e_{\mathrm{Ney}}^{*}(z) = \frac{\sqrt{v_1(z)}}{\sqrt{v_1(z)}+\sqrt{v_0(z)}}. \tag{11}$$

The total Objective I gain admits the telescoping identity

$$\begin{aligned}\mathcal{V}_{\mathrm{AIPW}}(\boldsymbol{e}_{\mathrm{Ney}}^{*}) - \mathcal{V}^{\mathrm{fs}}(\boldsymbol{e}_{\mathrm{I}}^{*,\mathrm{fs}}) = &\underbrace{\left\{\mathcal{V}_{\mathrm{AIPW}}(\boldsymbol{e}_{\mathrm{Ney}}^{*}) - \mathcal{V}^{\mathrm{zs}}(\boldsymbol{e}_{\mathrm{Ney}}^{*})\right\}}_{\text{Layer 1: LLM-enhanced efficiency gain}} + \underbrace{\left\{\mathcal{V}^{\mathrm{zs}}(\boldsymbol{e}_{\mathrm{Ney}}^{*}) - \mathcal{V}^{\mathrm{zs}}(\boldsymbol{e}_{\mathrm{I}}^{*,\mathrm{zs}})\right\}}_{\text{Layer 2: design-enhanced efficiency gain}} \\ &+ \underbrace{\left\{\mathcal{V}^{\mathrm{zs}}(\boldsymbol{e}_{\mathrm{I}}^{*,\mathrm{zs}}) - \mathcal{V}^{\mathrm{fs}}(\boldsymbol{e}_{\mathrm{I}}^{*,\mathrm{fs}})\right\}}_{\text{Layer 3: few-shot-enhanced efficiency gain}}.\end{aligned} \tag{12}$$

Layer 1 is calibrated LLM adjustment under fixed allocation, Layer 2 is the allocation gain from replacing classical Neyman allocation with the LLM-enhanced Neyman rule, and Layer 3 is the additional gain from few-shot aggregation. The first two are nonnegative oracle comparisons, while Layer 3 depends on prompt quality and need not be positive.

4.2 Welfare-enhanced benefits under Objective II

Objective II uses $\mathcal{L}_{\mathrm{F}}(\boldsymbol{e})$ as the objective and treats $\mathcal{V}(\boldsymbol{e}) \le C$ as a precision constraint. Calibrated LLM adjustment replaces $v_t(z)$ by $\tilde{v}_t(z) \le v_t(z)$ in the allocation-relevant criterion, so it can reduce $\mathcal{V}(\boldsymbol{e})$ for a fixed allocation and relax the feasible set $\{\boldsymbol{e} : \mathcal{V}(\boldsymbol{e}) \le C\}$. The welfare optimizer can then select more welfare-oriented allocations while satisfying the same precision budget.

This gives a three-layer Objective-II mechanism using the same layer names as in Objective I, but with a different interpretation. Layer 1, the LLM-enhanced efficiency gain, is a constraint-side effect: calibrated LLM adjustment reduces the allocation-relevant variance criterion and creates precision slack. Layer 2, the design-enhanced welfare gain, is the welfare-side benefit from re-optimizing $\mathcal{L}_{\mathrm{F}}(\boldsymbol{e})$ using this slack. Layer 3, the few-shot-enhanced welfare gain, is the additional welfare-side benefit when few-shot aggregation improves or stabilizes the auxiliary signal used in the constraint. Thus, unlike the Objective I decomposition, the Objective-II layers are mechanism components rather than an additive variance decomposition: Layer 1 is measured on the constraint side, whereas Layers 2–3 are measured on the welfare-objective side.

5 Theoretical investigation

This section gives large-sample guarantees for the proposed LLM-enhanced CARA procedure under predictable randomization. The proof has two components. Martingale arguments handle adaptive

treatment assignment, while a prompt-aggregation remainder controls the extra randomness from resampled few-shot demonstrations.

Assumption 5.1 (Regularity conditions)**.** The observed data $\{(W_i, Y_i(0), Y_i(1))\}_{i=1}^N$ are independently and identically distributed, where $W_i = (Z_i, X_i, U_i)$ and $Z_i \in \{1, \ldots, J\}$. For each $t \in \{0, 1\}$, $\mathbb{E}\{|Y_i(t)|^4\} < \infty$. There exists $\varepsilon > 0$ such that $\mathbb{V}\mathrm{ar}\{Y_i(t) \mid Z_i = z\} \geq \varepsilon$ for all z with $p(z) > 0$.

Assumption 5.2 (Sequential randomization)**.** For each stage s and stratum z, $e_s(z)$ is $\mathcal{H}_{s-1}$-measurable and satisfies $e_s(z) \in [\delta, 1 - \delta]$. Conditional on $(\mathcal{G}_{i-1}, Z_i)$, T_i is generated according to Eq. (1). In particular, $T_i \perp\!\!\!\perp \{Y_i(0), Y_i(1)\} \mid \mathcal{G}_{i-1}, Z_i$.

Assumption 5.3 (Nuisance estimation)**.** For each stage s and arm t, the estimators $\widehat{\mu}_{t,s}$, $\widehat{\mu}^F_{t,s}$, $\widehat{\mu}^\dagger_{t,s}$, $\widehat{\omega}_{t,s}$, and $\widehat{\tilde{v}}_{t,s}$ are consistent for μ_t, μ^F_t, $\mu^\dagger_t$, ω_t, and $\tilde{v}_t$ in L^2. The second-order remainder in the influence-function expansion is $o_p(N^{-1/2})$.

Assumption 5.4 (Few-shot LLM resampling stability)**.** For each subject i, arm t, and replicate $b = 1, \ldots, B$, the prompt-level output $Y_i^\dagger(t; \mathcal{S}_{i,t}^{(b)})$ is generated using only information available before prediction. The prompt-level outputs have uniformly bounded fourth moments. For each arm t, the aggregated signal admits the decomposition $Y_i^\dagger(t) = f_{\mathrm{FS},t}(W_i) + R_{i,t,B}$, where $f_{\mathrm{FS},t}$ is a stable limiting score and $N^{-1/2} \sum_{i=1}^N R_{i,t,B} = o_p(1)$.

Together, Assumptions 5.1–5.4 separate the main sources of randomness in the proposed design. Assumption 5.1 gives the moment and non-degeneracy conditions needed for martingale central limit arguments. Assumption 5.2 ensures predictable treatment assignment despite adaptive updates. Assumption 5.3 gives the nuisance-rate conditions for the influence-function expansion. Assumption 5.4 requires the few-shot aggregation remainder to be first-order negligible; uniformly bounded prompt-level variances together with $B^{-1/2} = o(N^{-1/2})$ provide one sufficient primitive condition.

Lemma 5.5 (Stability of few-shot aggregated signal)**.** *Under Assumption 5.4, the aggregated few-shot signal equals a stable limiting score plus a prompt-aggregation remainder whose contribution is first-order negligible at the $\sqrt{N}$ scale.*

Lemma 5.5 reduces the prompt-randomized few-shot signal to a stable auxiliary score plus a first-order negligible remainder.

Theorem 5.6 (Design strategy consistency)**.** *Under Assumptions 5.1–5.4, for each fixed stage $s = 2, \ldots, S$ and each stratum z with $p(z) > 0$,*

$$\widehat{e}_{s,\mathrm{I}}(z) \xrightarrow{\mathrm{p}} e^*_{\mathrm{I}}(z), \quad \widehat{e}_{s,\mathrm{II}}(z) \xrightarrow{\mathrm{p}} e^*_{\mathrm{II}}(z),$$

where $e^*_{\mathrm{I}}(z)$ and $e^*_{\mathrm{II}}(z)$ are the oracle allocation rules under Objectives I and II.

Theorem 5.6 justifies the stagewise plug-in allocation updates by showing convergence to the oracle rules.

Theorem 5.7 (Asymptotic normality)**.** *Under Assumptions 5.1–5.4, as $N \to \infty$,*

$$\sqrt{N}(\widehat{\tau} - \tau) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma^2_*), \quad \widehat{\mathbb{V}} \xrightarrow{\mathrm{p}} \sigma^2_*,$$

where σ^2_* is the full influence-function variance under the limiting predictable design.

Theorem 5.7 justifies inference under the adaptive design by using the full variance σ^2_*, while $\mathcal{V}(\boldsymbol{e})$ is only the allocation-sensitive component used to derive the design rules.

Theorem 5.8 (Efficiency gain)**.** *Under Assumptions 5.1–5.4, the allocation-relevant criterion satisfies two oracle comparisons. For any feasible CARA policy $\boldsymbol{e}$,*

$$\mathcal{V}(\boldsymbol{e}) \leq \mathcal{V}_{\mathrm{AIPW}}(\boldsymbol{e}).$$

Moreover, if $\boldsymbol{e}^*_{\mathrm{I}}$ is the Objective I oracle policy and $\boldsymbol{e}^*_{\mathrm{Ney}}$ is the classical Neyman policy in Eq. (11)*, then*

$$\mathcal{V}(\boldsymbol{e}^*_{\mathrm{I}}) \leq \mathcal{V}(\boldsymbol{e}^*_{\mathrm{Ney}}).$$

Theorem 5.8 formalizes the nonnegative oracle parts of the Objective-I decomposition, namely calibrated residual-variance reduction and allocation using effective residual variances.

Theorem 5.9 (Asymptotic equivalence under few-shot aggregation)**.** *Assume Assumptions 5.1–5.4 hold. Let $\widehat{\tau}$ be the proposed estimator using the aggregated few-shot signal $Y_i^\dagger(t)$, and let $\widehat{\tau}^*$ be the same estimator with $Y_i^\dagger(t)$ replaced by $f_{\mathrm{FS},t}(W_i)$ everywhere. Then*

$$\widehat{\tau} - \widehat{\tau}^* = o_p(N^{-1/2}).$$

Thus, $\widehat{\tau}$ has the same first-order influence function and asymptotic distribution as $\widehat{\tau}^*$.

Theorem 5.9 formalizes the few-shot layer: under resampling stability, prompt randomness does not affect first-order inference, although finite-sample gains still depend on prompt quality.

6 Simulation studies

We evaluate the proposed LLM-enhanced CARA procedure on synthetic data calibrated to the BRIGHTEN study [Pratap et al., 2022], a fully remote smartphone-based randomized trial for depression management with structured baseline covariates and free-text surveys.

6.1 Simulation setup

BRIGHTEN synthetic data generation. BRIGHTEN enrolled adults with clinically significant depressive symptoms and randomized participants to app-delivered interventions, with follow-up over 12 weeks. We use structured baseline covariates X_i, two baseline free-text survey fields U_i, and the PHQ-9 depression severity score as the continuous outcome Y_i.

We use a BRIGHTEN-calibrated synthetic superpopulation of size 20,000 with structured covariates, recovered baseline free-text fields, and arm-specific PHQ-9 potential outcomes. The DGP preserves the original BRIGHTEN-calibrated conditional outcome structure and adds an orthogonal text-augmentation component derived from LLM auxiliary predictions. The arm-specific outcomes are recentered to preserve the original BRIGHTEN marginal means and clipped to the valid PHQ-9 range $[0, 27]$. The conditional mean models incorporate both structured covariates and text-derived components so that the LLM auxiliary signal has nontrivial but imperfect predictive content for the generated outcomes.

For LLM prompting on synthetic subjects, we recover each subject's pre-treatment text by matching the synthetic text representation to semantically similar observed BRIGHTEN free-text responses. We pair the recovered snippets with the synthetic structured covariates to form a complete baseline record. For each subject and arm $t \in \{0, 1\}$, Qwen3-8B receives a fixed structured prompt that serializes the baseline record and requests a continuous PHQ-9 prediction under arm t. The zero-shot prompt contains only the target subject's baseline record. The few-shot prompt also includes arm-specific demonstrations sampled from accrued trial data, and predictions are averaged over B resampled prompt sets as described in Section 2.3. The reported few-shot Qwen3-8B auxiliary predictions use $B = 4$ resampled demonstration sets per subject-arm pair.

Design setup. We run $R = 300$ Monte Carlo replicates. In each replicate, we sample $N = n_1 + n_2$ units without replacement from the superpopulation, set $n_1 = n_2$, and randomly permute enrollment

Figure 1: Efficiency gains based on cross-fitted influence-function variance estimates. MISSING denotes an unreported education category.

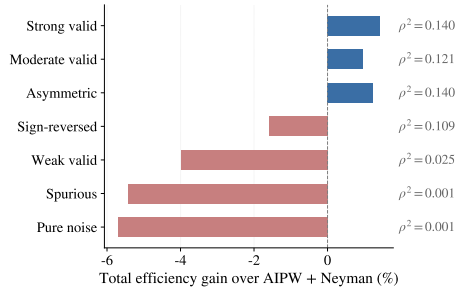

Figure 2: Signal scenarios vary auxiliary-signal strength, direction, and arm-specific reliability; construction details are given in Appendix J.

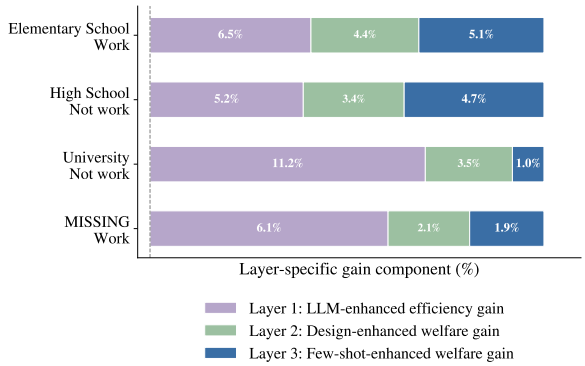

Figure 3: Objective-II mechanism components. Layer 1 is an LLM-enhanced efficiency gain that relaxes the precision constraint; Layers 2–3 are welfare-side failure-rate reductions. The bars are not an additive welfare decomposition.

order. Stage 1 uses equal randomization, $e_1(z) \equiv 1/2$. Stage 2 uses stratum-specific allocation probabilities computed from stage-1 estimates and truncated to $[\delta, 1-\delta]$.

Auxiliary outcomes are arm-specific PHQ-9 predictions generated from baseline information. We compare non-LLM text baselines based on TF-IDF or sentence embeddings with random forests or ridge regression, zero-shot Qwen3-8B prompting, and few-shot Qwen3-8B prompting. For trained non-LLM baselines and few-shot prompting, we use 2-fold cross-fitting so that each target prediction uses a model or demonstration pool built outside the target fold.

Evaluation metrics. For Objective I, we report efficiency gains as percent reductions in the average cross-fitted influence-function variance relative to AIPW under the classical Neyman allocation, decomposed into estimation, allocation, and few-shot components. For Objective II, we report reduction in the adverse PHQ-9 event rate, where the event is defined as PHQ-9 ≥ 15, corresponding to the two most severe PHQ-9 categories, relative to the Ethical design. Unless otherwise stated, the main experiments use the calibrated procedure without the validation-gated safeguard. Real-data-calibrated synthetic experiments are commonly used in CARA methodology because counterfactual outcomes and repeated adaptive trials are not observed in a single realized study [Hu et al., 2015, Villar and Rosenberger, 2018, Zhu and Zhu, 2023].

6.2 Simulation results

Figure 1 compares cross-fitted efficiency gains across auxiliary prediction methods. Few-shot LLM aggregation yields the largest gains across the displayed strata. Zero-shot prompting gives smaller but consistent improvements, while non-LLM text baselines give only modest gains. Figure 2 shows that these gains depend on signal reliability. Strong, moderate, and asymmetric valid signals produce positive gains. By contrast, sign-reversed, weak, spurious, and pure-noise signals reduce efficiency relative to AIPW with Neyman allocation.

Figure 3 summarizes the three-layer Objective-II mechanism across education and work-status strata. Layer 1 creates precision slack through LLM-enhanced efficiency. Layers 2 and 3 use the design-enhanced and few-shot-enhanced welfare components to convert that slack into lower adverse outcome rates. The bars represent mechanism components rather than an additive decomposition. Layer 1 is measured on the constraint side, whereas Layers 2 and 3 are measured on the welfare-objective side. Although the Objective-I allocation gain is modest in this BRIGHTEN-calibrated setting, the adaptive allocation component remains central under Objective II. It converts LLM-enhanced precision slack into welfare-oriented assignment changes and lower adverse outcome rates. Appendix J reports additional diagnostics on safeguards, screening, and finite-B behavior.

7 Discussion

We developed an LLM-enhanced CARA framework that uses calibrated arm-specific auxiliary predictions to guide adaptive allocation and final inference. The framework preserves design-known

randomization and supports cross-fitted inference. It treats LLM outputs as auxiliary signals rather than plug-in counterfactual outcomes. Calibration and effective residual variances then link those signals to both estimation and allocation.

The main limitations concern finite-sample stability and auxiliary-signal reliability. Sparse strata, noisy outcomes, prompt variation, population shift, and query cost can reduce the value of LLM signals. The method may improve statistical efficiency and participant welfare when auxiliary signals are reliable. However, unreliable or miscalibrated signals can steer allocation away from its intended target. Applications should therefore use validation diagnostics, truncation, safeguards, and finite-B stability checks. Future work will develop sharper stabilization rules and diagnostics for unreliable auxiliary predictions.

References

Aurélien Bibaut and Nathan Kallus. Demystifying inference after adaptive experiments. *Annual Review of Statistics and its Application*, 12(1):407–423, 2025.

Jessica Dai, Paula Gradu, and Christopher Harshaw. Clip-ogd: An experimental design for adaptive neyman allocation in sequential experiments. *Advances in Neural Information Processing Systems*, 36:32235–32269, 2023.

Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–331, 1998.

Jinyong Hahn, Keisuke Hirano, and Dean Karlan. Adaptive experimental design using the propensity score. *Journal of Business & Economic Statistics*, 29(1):96–108, 2011.

Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

Feifang Hu and William F Rosenberger. *The Theory of Response-adaptive Randomization in Clinical Trials*. John Wiley & Sons, New York, NY, 2006.

Jianhua Hu, Hongjian Zhu, and Feifang Hu. A unified family of covariate-adjusted response-adaptive designs based on efficiency and ethics. *Journal of the American Statistical Association*, 110(509): 357–367, 2015.

Yanqing Hu and Feifang Hu. Asymptotic properties of covariate-adaptive randomization. *Annals of Statistics*, 40(3):1794–1815, 2012.

Xinwei Ma and Jingshen Wang. Robust inference using inverse probability weighting. *Journal of the American Statistical Association*, 115(532):1851–1860, 2020.

Ojash Neopane, Aaditya Ramdas, and Aarti Singh. Logarithmic neyman regret for adaptive estimation of the average treatment effect. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan, editors, *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 4303–4311. PMLR, 03–05 May 2025. URL `https://proceedings.mlr.press/v258/neopane25a.html`.

Jerzy S Neyman. On the application of probability theory to agricultural experiments. essay on principles. section 9.(translated and edited by dm dabrowska and tp speed, statistical science (1990), 5, 465-480). *Annals of Agricultural Sciences*, 10:1–51, 1923.

Molly Offer-Westort, Alexander Coppock, and Donald P Green. Adaptive experimental design: Prospects and applications in political science. *American Journal of Political Science*, 65(4): 826–844, 2021.

Abhishek Pratap, Ava Homiar, Luke Waninger, Calvin Herd, Christine Suver, Joshua Volponi, Joaquin A Anguera, and Pat Areán. Real-world behavioral dataset from two fully remote smartphone-based randomized clinical trials for depression. *Scientific Data*, 9(1):522, 2022.

David S Robertson, Kim May Lee, Boryana C López-Kolkovska, and Sofía S Villar. Response-adaptive randomization in clinical trials: From myths to practical considerations. *Statistical Science*, 38(2):185, 2023.

William F Rosenberger and John M Lachin. *Randomization in clinical trials: theory and practice*. John Wiley & Sons, 2015.

William F Rosenberger, Nigel Stallard, Anastasia Ivanova, Cherice N Harper, and Michelle L Ricks. Optimal adaptive designs for binary response trials. *Biometrics*, 57(3):909–913, 2001.

Xinrui Ruan, Xinwei Ma, Yingfei Wang, Waverly Wei, and Jingshen Wang. Can language models boost the power of randomized experiments without statistical bias? *arXiv preprint arXiv:2510.05545*, 2025.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Anastasios A Tsiatis. *Semiparametric Theory and Missing Data*. Springer, New York, NY, 2006.

Sofía S Villar and William F Rosenberger. Covariate-adjusted response-adaptive randomization for multi-arm clinical trials. *Biometrics*, 74(3):839–848, 2018.

Li-Xin Zhang, Feifang Hu, Siu Hung Cheung, and Wai Sum Chan. Asymptotic properties of covariate-adjusted response-adaptive designs. *Annals of Statistics*, 35(1):1166–1182, 2007.

Jinglong Zhao. Adaptive Neyman allocation. arXiv preprint arXiv:2309.08808, 2023.

Hai Zhu and Hongjian Zhu. Covariate-adjusted response-adaptive designs based on semiparametric approaches. *Biometrics*, 79(4):2895–2906, 2023.